



AVALIAÇÃO DE ABORDAGENS PROBABILÍSTICAS DE EXTRAÇÃO DE TÓPICOS EM DOCUMENTOS CURTOS

Michel Chagas da Costa (apresentador)¹

Dr. Denio Duarte²

Resumo: Nas últimas décadas, os textos curtos prevalecem na Internet, e principalmente com o crescimento das mídias sociais, os textos curtos passaram a predominar na Web. Milhões de usuários geram um grande número de textos curtos todos os dias. Com este grande volume de dados disponíveis, é desejável que o aprendizado de máquina extraia tópicos a partir de coleções de documentos curtos, visando auxiliar as mais diversas aplicações, como descobrir o perfil de interesse do usuário, sugerir buscas ou organização destes dados. Algumas abordagens probabilísticas convencionais de modelagem de tópicos, como o LDA (Latent Dirichlet Allocation), provaram ter sucesso ao extrair tópicos em documentos longos, porém se mostraram limitadas ao extrair tópicos em documentos curtos. Notou-se problemas de dados esparsos e desbalanceados em coleções de documentos curtos, pois em modelos de tópicos baseados no LDA, o número de ocorrência de uma palavra em um documento é fundamental para que se possa inferir tópicos. Novas abordagens, com diferentes adaptações, foram surgindo, com o objetivo de extrair tópicos em coleções de documentos curtos. Este trabalho se propõe a avaliar o uso de cinco destas abordagens, com características distintas e que propõem uma solução para o problema de dados esparsos em coleções de documentos curtos, utilizando métricas e coleções de documentos com o número médio de palavras sendo diferenciado em cada uma destas coleções. Criar grandes documentos a partir de pequenos documentos, enriquecer um documento curto através de documentos longos externos e considerar a coleção de documentos curtos como uma rede de palavras, são algumas técnicas utilizadas por estas novas abordagens em modelagem de tópico que visam tratar o problema dos dados esparsos ao inferir tópicos em documentos curtos.

Palavras-chave: Aprendizado de máquina. Modelagem de tópicos. Textos curtos. Avaliação. LDA.

¹ Graduando, Universidade Federal da Fronteira Sul, *campus* Chapecó, contato: costa.michel10@hotmail.com

² Doutor, Universidade Federal da Fronteira Sul, *campus* Chapecó, contato: duarte@uffs.edu.br



Anais do SEPE – Seminário de Ensino, Pesquisa e Extensão
Vol. VIII (2018) – ISSN 2317-7489



Categoria: Pesquisa

Área do Conhecimento: Ciências Exatas e da Terra

Formato: Comunicação Oral